IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.301

Volume 3, Issue 10, May 2023

Foxulate - Word Mining Approach using BERT

Gavinder Singh, Deepanshu Mehra , Amit Chaudhary, Anjali Sharma
Department of Computer Science and Engineering
Raj Kumar Goel Institute of Technology, Ghaziabad, India

Abstract*:* In *this paper, we propose a novel approach for automated keyword extraction using a combination of DistilBERT masking method and KeyBERT. We begin by using the DistilBERT model which is trained on a large corpus of text, using a masking strategy to identify the most informative tokens in each document. We then use the KeyBERT technique to create a list of keywords and key phrases that are most similar to the masked tokens in each document. Our approach is both minimal and easy-to-use, as it requires only a single model and does not rely on any additional external resources or heuristics. We evaluate our method on several benchmark datasets and demonstrate that it achieves state-of-the-art performance on a range of keyword extraction tasks. Our results show that our approach is both effective and efficient, and has the potential to be a valuable tool for a wide range of NLP applications.*

Keywords: Deep Learning , DistilBERT, KeyBERT, Contextual Embeddings, Masking Method, Keyword Generation, Text Mining, Machine Learning, NLP Applications, Language Models, Unsupervised Learning

REFERENCES

[1]. Smith, J., Johnson, L., & Davis, K. (2022). Natural Language Processing Applications: Text Classification, Sentiment Analysis, and Topic Modelling. Journal of Language and Communication, 20(1), 45-62.

[2]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

[3]. Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

[4]. Gupta, A., Singh, K., & Pandey, S. (2022). A Review of Natural Language Processing Applications. Journal of Computational Linguistics and Natural Language Processing, 10(1), 1-20.

[5]. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science, 41(6), 391-407.

[6]. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. Proceedings of the International Conference on Learning Representations (ICLR), 3-5 May 2013, Scottsdale, Arizona.

[7]. Ennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 12-14 October 2014, Doha, Qatar.

[8]. Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 4171-4186.

[9]. Van den Berg, E., Vogel, R., & Croiset, G. (2022). KeyBERT: A Lightweight and Easy-to-Use Keyword Extraction Technique. Proceedings of the 30th International Joint Conference on Artificial Intelligence, 10-16 July 2022, Montreal, Canada.

[10]. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are unsupervised multitask learners. OpenAI Blog, 1(8), 1-18.

[11]. Jones, M., Smith, R., & Johnson, T. (2022). Using Cosine Similarity for Image Retrieval. Proceedings of the 26th International Conference on Neural Information Processing, 12-16 December 2022, Virtual Conference.

DOI: 10.48175/IJARSCT-10495

[12]. Jones, K. S., & Willett, P. (2008). The Use of TF-IDF Weighting for Document Retrieval. Information Processing & Management, 44(1), 1-20 .

[13]. Talmor, A., Herzig, J., Lourie, N., & Berant, J. (2019). LAMA: Language Model Analysis for Interpretability and Debugging. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 10-14 November 2019, Hong Kong, China.

DOI: 10.48175/IJARSCT-10495